

Characterization of South American Educational Web Domains

Fernando R. A. Bordignon

Laboratorio de Redes de Datos
Departamento de Ciencias Básicas, Universidad Nacional de Luján
Luján, B. A., Argentina
bordi@unlu.edu.ar

Gabriel H. Tolosa

Laboratorio de Redes de Datos
Departamento de Ciencias Básicas, Universidad Nacional de Luján
Luján, B. A., Argentina
tolosoft@unlu.edu.ar

Abstract

The aim of this work is to characterize the educational web domains of Argentina, Bolivia, Chile, Paraguay, Perú and Uruguay. For this purpose, a sample of each information domain was collected using an automatic crawling tool. These domains were later analyzed according to different aspects of their contents, links and technologies in order to determine differences and similarities from their most representative parameters.

From this study we can derive some interesting facts: a) the amount of contents in all the countries has increased over the last year, b) it is necessary to increase the ranking of pages in educational domains, especially by providing incoming links from intra-domains, as we found a high percentage of pages with zero in-degree, and c) new technologies are needed to manage dynamic pages within educational institutions.

Keywords: Academics Webs Spaces, South American Educational Institutions, Characterization.

Resumen

El presente trabajo tiene por objetivo caracterizar los espacios webs educativos de Argentina, Bolivia, Chile, Paraguay, Perú y Uruguay. Se realizó la recolección de una muestra de los distintos espacios de información utilizando un herramienta de recolección automática. Luego, se analizaron aspectos de cada dominio en lo referido a contenidos, enlaces y tecnologías utilizadas, de manera de poder establecer similitudes y diferencias en cuanto a los parámetros más representativos.

Del estudio surgen las siguientes observaciones: a) la cantidad de contenidos de todos los países participantes del estudio se ha incrementado considerablemente el último año, b) es necesario que las páginas de los dominios educativos tengan una mayor visibilidad, especialmente entre sus pares. Se ha encontrado para todos los países un importante porcentaje de páginas con grado entrante igual a cero y c) aún falta incorporar tecnología que permita administrar páginas dinámicas en las instituciones educativas.

Palabras clave: Espacios Webs Educativos, Instituciones Educativas Sudamericanas, Caracterización.

1 Introduction

The study of the different characteristics of the Web arouses great interest nowadays, mainly because of the need to understand and model its structure and evolution. This understanding makes it possible to design better strategies to find and collect information, tune up ranking algorithms and create high quality websites and applications that make use of the Web as an information repository [11].

Collecting data from the web space is a complex task which requires the use of large-scale computational resources given its size and distribution. The following studies were performed using samples of different size and domains, and applying multiple collection strategies. More specifically, we examined national domains [2, 4, 5, 13, 16], as they constitute a very interesting study set due to their good balance between diversity and completeness [6].

Although these studies provide valuable information, we found it more useful to focus on specific portions of the selected sample, which have their own characteristics. In this paper, we carried out a study of the national educational domains of South American countries, which exhibit their own distinctive qualities given the fact that they are built and maintained by a particular community and mostly used by specific groups. Under this assumption, and considering the constant growth and evolution of web spaces in their size, technologies and services, it becomes necessary to gain a deeper insight into their nature. From a social perspective, we also tried to determine the patterns and trends that could enable the design of better strategies to facilitate the access to the information contained in these domains.

This work comprises studies of the educational domains of Argentina, Chile, Uruguay, Paraguay, Perú and Bolivia, and aims at characterizing their current state in order to discover any differences and similarities from their most representative parameters. The results obtained might serve as a starting point for studying the evolution of these domains and for helping people in charge of them make decisions about their development. To our knowledge, there are no similar studies focused exclusively on educational domains. A first approach, albeit on a smaller scale, can be found in [19], which contains link-level analyses of Argentine universities. It also determined the distribution of these organizations, their degree of connectivity and rankings, among other link-level state indicators. Our research, however, is based on works describing the national domains of Chile [4], Brazil [16] and Spain [5].

2 Data Collection

The data for our study was collected using a *crawler* called WIRE [10]. This software works in automatic mode and is specially designed to handle large volumes of data. The collection process was conducted during November and December 2005 using a PC server with a 2.4 GHz processor, 1 GB of RAM and running the Linux Debian Sarge operating system.

The software was configured to limit the maximum crawling depth to 15 for static pages and 5 for dynamic pages. Only web pages were collected, with a maximum of 25,000 per site. Although this limitation was set to 5,000 in other studies [4], for this work we assumed that educational spaces would be richer in content and, being a restricted domain, they could be exhaustively collected, thus improving the quality of the sample.

We defined the educational web of a country as the set of websites that were under the .edu second level domain (for example, .edu.ar for Argentina). We also included many addresses of education-related institutions such as ministries, libraries and portals. For each country, the collection process started from a set of initial addresses taken from Google¹ and Netcraft². Chile represents an exception as it does not use the .edu domain. In this case, the crawler used a list of URLs of institutions which were supposedly educational, duly verified by a human assistant. Table 1 shows the sizes of the resulting collections. It is easy to notice how they vary in volume, although in each case the crawling process stopped once the program was unable to find any new URL to visit. The first noticeable fact is the small amount of duplicate pages found.

	edu.ar		edu.bo		(edu).cl		edu.pe		edu.py		edu.uy	
Pages	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
Total	375,496		22,293		147,152		165,237		7,208		70,896	
Unique	370,833	98.76	22,233	99.73	146,431	99.51	163,892	99.19	7,190	99.75	70,196	99.01
Duplicates	4,663	1.24	60	0.27	721	0.49	1,345	0.81	18	0.25	700	0.99
Static	329,981	87.88	15,591	69.94	116,111	78.91	118,905	71.96	5,957	82.64	59,590	84.05
Dynamic	45,515	12.12	6,702	30.06	31,041	21.09	46,332	28.04	1,251	17.36	11,306	15.95

Table 1 – Collection sizes

For the analysis and presentation of the results, we followed the methodology proposed in [7], grouping features according to contents, links and technologies. In each case, we also provide data at different levels of granularity.

3 Contents

3.1 Page Size

The size of the pages retrieved from each educational domain was analyzed. Such distribution can be represented by a power law, where there are few big-sized pages and many small pages. For the collection phase, the software was set to download up to 400 Kb per page. Figure 1 shows graphs of the size distributions. The lines of fit (power laws of parameter β) were estimated over the center of each distribution. In the studies taken as reference, an exponent $\beta = 3.2$ was observed for Chile [4], whereas a value of $\beta = 3.5$ was found for Brazil [16] and $\beta = 2.25$ for Spain [5].

3.2 Page Age

To obtain the age of each downloaded page, we used the “last-modified” value in the HTTP header sent by the server to the crawler. All dates before the creation of the Web or referring to the future were rejected. Here, it should be noted that dynamic pages, which are created at the time of the request, are generally 0 years old. For the distributions, the values of the number of pages created in the last year were adjusted according to the proportion of static pages found. Figure 2 displays these results. It should be noted that all countries have grown considerably over the last year whereupon their educational webs should be considered young. This matches the continuous growth of the web, which explains why there are more new sites than old ones [17]. In the study of the Chilean web [4], a 25% of new pages was observed nationwide during 2004 (the last analyzed year). However,

¹ <http://www.google.com/>

² <http://www.netcraft.com/>

educational domains exhibit a growth or update which is higher than that of previous years and which varies roughly between 37% and 70%.

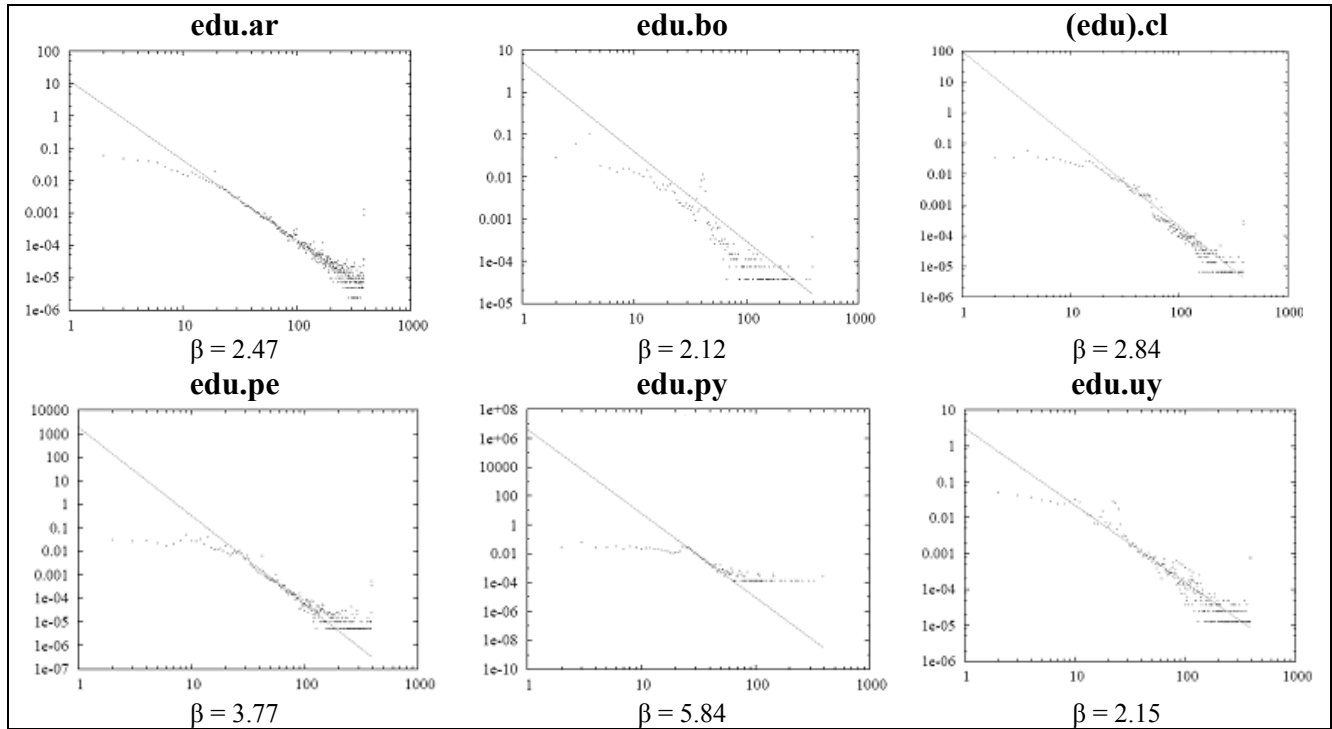


Figure 1 – Page age distributions. The X-axis represents the page size in Kb while the Y-axis indicates their relative frequency. Power-law parameter β is given in each case.

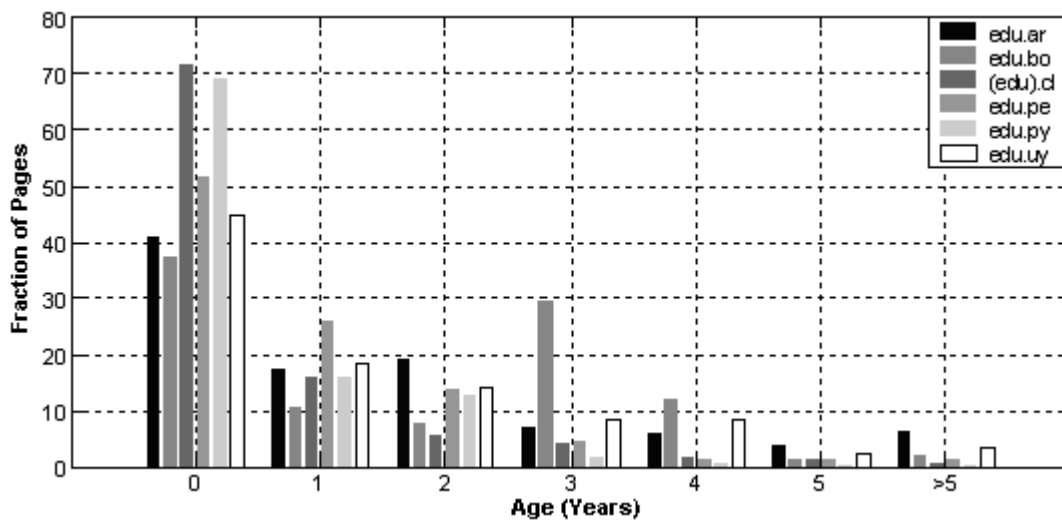


Figure 2 – Page age distribution.

3.3 Language

A text sample from the collected pages was extracted for language recognition. The proportions of the main languages recognized are presented in Table 2.

	edu.ar	edu.bo	(edu).cl	edu.pe	edu.py	edu.uy
Spanish	58.03	51.85	78.40	69.66	90.91	67.65
English	39.90	48.15	20.80	28.46	9.09	25.49
Other	2.07	0.00	0.80	1.88	0.00	6.86

Table 2 – Language proportions within the studied domains

As we had expected, Spanish is the predominant language in all cases, since it is the official language of all the studied countries. However, in both Argentina and Bolivia there is a high percentage of English content, which amounts to 40% and 48% respectively. After a manual inspection, we found copies (*mirrors*) of large English sites within the Argentinian web. Such an example is the website of “*The European Mathematical Information Service (EMIS)*”, whose copies are kept by two national universities: <http://www.emis.unne.edu.ar/> and <http://www.emis.famaf.unc.edu.ar/>. In the second case, 25,000 documents were registered, as limited by the maximum number of downloadable pages per site in the *crawler*. Inside the Argentinian sample, there are also some private secondary schools whose websites are available in both English and Spanish, such as St John’s School (<http://www.stjohnsschool.edu.ar/>) and St Brendan’s College (<http://www.stbrendans.esc.edu.ar/>). In Bolivia, we also found some schools with English websites like Cochabamba Cooperative School (<http://www.ccs.edu.bo/>). Although we did not found large mirrored sites in English, the small size of the Bolivian web emphasizes this proportion.

3.4 Number of Websites and Pages per Site

This section describes the results shown in Table 3 about the number of successfully downloaded websites and the average number of pages within each site.

	Number of Sites	Average Number of Pages
edu.ar	1964	211
edu.bo	218	122
(edu).cl	1967	76
edu.pe	704	283
edu.py	118	62
edu.uy	289	274

Table 3 – Number of Sites and Average Number of Pages

There is a significant difference in almost every domain with respect to the same values in national webs. For example, there are 53 pages per site in Chile [4], 85 in Brazil [16] and 52 in Spain [5]. A possible explanation for this phenomenon lies in the way domain names were obtained. While a commercial domain could be granted to or obtained by any person or company, the acquisition of educational domains is subject to many conditions as they are reserved for institutions, which is why their websites are often larger. The only exception, as in the rest of this study, is the educational web of Paraguay, which was found to be the smallest.

3.5 Size of Websites and their Pages

The average size in bytes of the websites and their pages was studied considering only HTML pages, i.e. discarding images and other objects. Website sizes are uneven, although the average size per page is comparable in most cases (except for Bolivia). Table 4 contains the calculated averages for sites and pages.

	Average Size per Site (Mb)	Average Size per Page (Kb)
edu.ar	1.85	8.98
edu.bo	0.65	5.45
(edu).cl	0.73	9.81
edu.pe	2.60	9.40
edu.py	0.57	9.41
edu.uy	2.47	9.23

Table 4 – Average Size of Sites and Pages per Site

4 Links

The web can be modeled as a directed graph (the webgraph), where nodes represent pages and edges are links. Broder [9] studied the basic features of the webgraph and found a property which states that the distribution of the vertices' in-degree follows a power law, something very common in scale-free networks [14, 8], which have an uneven distribution of nodes and links [1]. On the other hand, Dill [12] observed that the web shows invariant characteristics, that is, that the properties found in particular portions are consistent with those of the whole network.

4.1 In-Degree

We first studied in-degree, which represents the number of links received from other pages. This feature becomes important since it is used by search engines as an element to aid the computation of rankings for their final results lists. Typically, degree distributions are highly skewed given that there are numerous pages with few links and only few pages with a large number of links, a distribution which can be modeled by a power law. On this aspect, Kleinberg defined the concepts of Hub and Authoritative pages [15] for his HITS algorithm. Alternatively, PageRank [18] is the other classic algorithm that defines the importance or popularity of a page based on its in-degree. Figure 3 shows the in-degree distributions for the studied domains.

In Broder's study using a large-scale web sample [9], an exponent $\beta = 2.1$ was observed for the power law describing the in-degree distribution, while Brazil [16] was found to have $\beta = 1.0$, Chile [4] $\beta = 2.0$ and Spain $\beta = 2.1$ [5].

Interestingly, we found a 33% of pages with zero in-degree in edu.ar, 54% in edu.bo, 20% in (edu).cl, 34% in edu.pe, 23% in edu.py and 39% in edu.uy. These figures are of great importance if we consider how ranking algorithms make use of this information to determine popularity.

4.2 Out-Degree

To complete the study, we also analyzed out-degree, which is the number of links from a single page to other pages. Out-degree distributions are consistent with an imperfect power law, especially for low values (Figure 4).

These distributions portray different scenarios, as seen in Bolivia and Paraguay, possibly due to the small size of their educational web spaces. In Broder's study [9], a power law of parameter $\beta = 2.72$ was found, while $\beta = 2.6$ was observed for Chile [4], $\beta = 2.7$ for Brazil [16] and $\beta = 4.2$ for Spain [5].

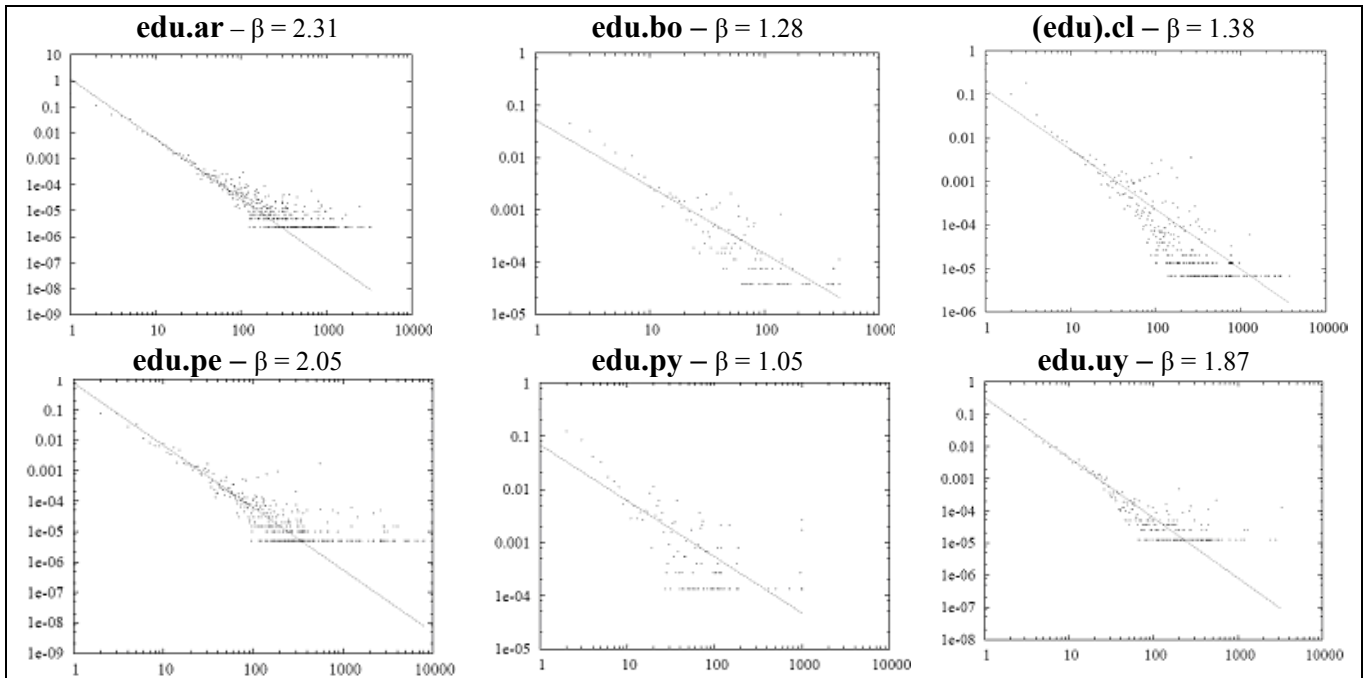


Figure 3 – In-degree distributions. The X-axis shows the number of in-links while the Y-axis represents the fraction of pages. Power-law parameter β is given in each case.

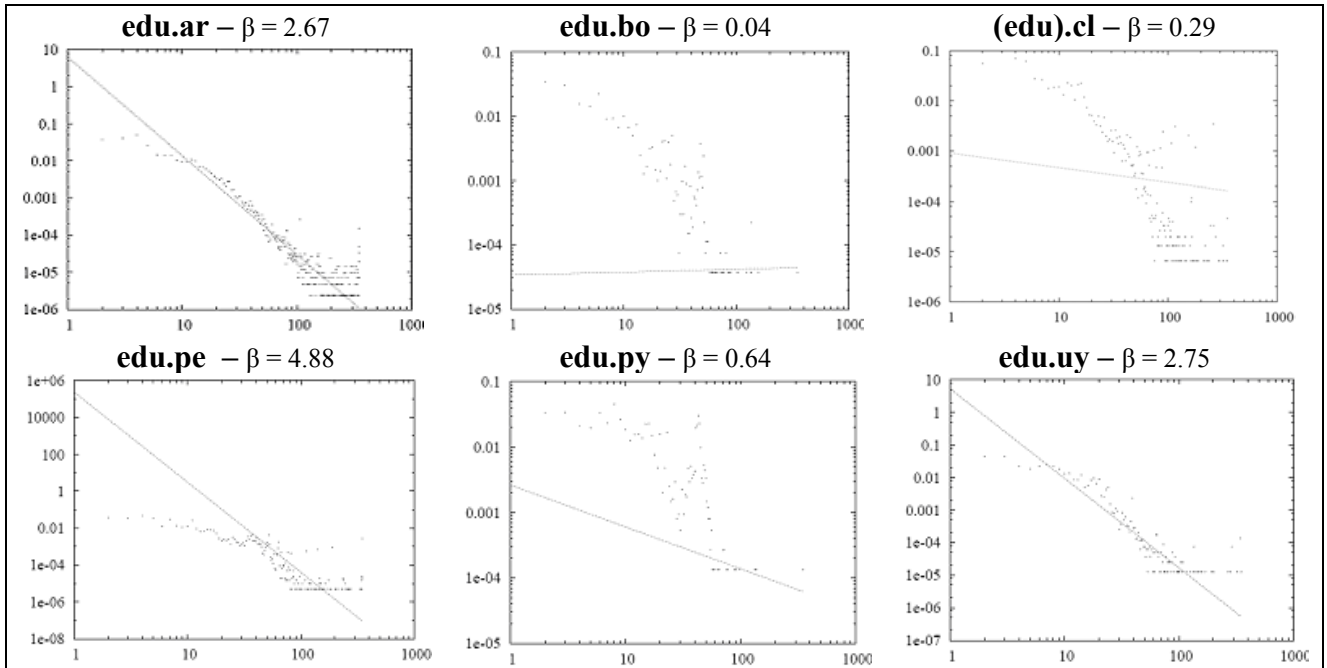


Figure 4 – Out-degree distributions. The X-axis shows the number of out-links while the Y-axis represents the fraction of pages. Power-law parameter β is given in each case.

4.3 Ranking of Pages and Websites

The ranking of pages within each educational domain is also an important matter that deserves attention. As was mentioned earlier, one of the most well-known algorithms is PageRank, which uses information of incoming links to determine the importance of a webpage. PageRank values were computed for each analyzed page, yielding distributions which were also modeled by power laws. In this case, we observed an exponent $\beta = 1.86$ in edu.ar, (edu).cl, edu.pe and edu.uy, $\beta = 1.81$ in edu.bo and $\beta = 1.75$ in edu.py.

Next, we grouped the pages into sites to obtain their rankings. In the following table (Table 5), we include the first 15 sites under each domain sorted by the sum of the PageRank values of their individual pages.

edu.ar	edu.bo	(edu).cl
g.unsa.edu.ar	www.ceub.edu.bo	www.eclac.cl
weblog.mendoza.edu.ar	www.lincoln.edu.bo	www.udp.cl
www.fcen.uba.ar	www.umss.edu.bo	www.notredame.cl
web.fcen.uba.ar	www.uto.edu.bo	www.colegionotredame.cl
pub.ufasta.edu.ar	www.fiumsa.edu.bo	www.colegioapoquindo.cl
www.emis.famaf.unc.edu.ar	www.uapnet.edu.bo	elqui.dsc.utfsm.cl
beba.fisica.unlp.edu.ar	www.unikuljis.edu.bo	www.iaemaristas.cl
www.unsa.edu.ar	www.unslp.edu.bo	www.salesianoslinares.cl
www.rec.uba.ar	www.loyola.edu.bo	www.ejbiotechnology.cl
www.salvador.edu.ar	www.uajms.edu.bo	www.isfmaristas.cl
www.agro.uba.ar	www.ucbcba.edu.bo	www.colegiosanagustindeatacama.cl
softwarelibre.unsa.edu.ar	www.colegiocise.edu.bo	www.colegio-bautista.cl
www.cema.edu.ar	www.bv.umsanet.edu.bo	www.colegiosalesiano.cl
bo.unsa.edu.ar	orion.ucb.edu.bo	www.derechoshumanos.cl
www.palermo.edu.ar	www.upb.edu	www.liceoaleman.cl

edu.pe	edu.py	edu.uy
www.champagnat.edu.pe	www.quanta.net.py	www.rau.edu.uy
blog.pucp.edu.pe	cdu.cnc.una.py	www.dsm.edu.uy
www.amersol.edu.pe	www.una.py	www.anep.edu.uy
www.unsa.edu.pe	newton.cnc.una.py	www.reu.edu.uy
photos.amersol.edu.pe	www.webescuela.edu.py	www.montevi.edu.uy
www.bibliocentral.udep.edu.pe	www.aulaviva.edu.py	www.ucu.edu.uy
sisbib.unmsm.edu.pe	www.campoalto.edu.py	www.ccee.edu.uy
www.chanel.edu.pe	www.ing.una.py	www.safa.edu.uy
notes.uni.edu.pe	www.uninet.com.py	www.fing.edu.uy
www.derecho.usmp.edu.pe	www.uaa.edu.py	www.cmat.edu.uy
www.lamolina.edu.pe	www.fpuna.edu.py	www.mec.gub.uy
www.udep.edu.pe	www.rec.una.py	ie.fing.edu.uy
www.unmsm.edu.pe	www.falevidelguap.edu.py	www.seminario.edu.uy
www.upch.edu.pe	www.cnc.una.py	www.lospilares.edu.uy
info.upc.edu.pe	www.arq.una.py	www.memfod.edu.uy

Table 5 – First 15 sites under each domain sorted by the sum of their PageRank values.

5 Technologies

5.1 Downloaded Pages

The aim of this first study is to obtain the number of correctly downloaded pages from the total URLs analyzed. Normally, a web client (the WIRE *crawler* in our case) opens a TCP connection to the corresponding web server and requests the desired resource through the HTTP protocol. The server responds with a state code, whose evaluation is useful for determining whether the page was downloaded correctly or the reasons why it could not be retrieved. For clarity, HTTP response codes were grouped into four categories (Table 6).

Resultado	Descripción	Códigos HTTP incluidos
OK	Successful download	OK (200), PARTIAL (206)
MOVED	Redirection to another page	MOVED (301), FOUND (302), TEMPORARY REDIRECT (307)
SERVER ERROR	Server-side errors	INTERNAL SERVER ERROR (500), NO CONTENT (504), BAD GATEWAY (502), UNAVAILABLE (503)
FORBIDDEN	Forbidden requests	UNAUTHORIZED (401), FORBIDDEN (403), NOT ACCEPTABLE (406)

Table 6 – Groups of HTTP codes

Figure 5 displays the results of download attempts. Clearly, percentages of successful downloads range from 78% in edu.uy to 95% in (edu).cl. Baeza-Yates observed 86% of successful downloads in his study of the Chilean web [4] while Modesto found a similar value in Brazil [16]. On the other hand, redirections account for a significant portion in edu.ar (6%) and edu.uy (4%) but not in the rest of the domains. Such redirections appear when a website is being reorganized or changes are being made to their names or the location of their pages. Finally, the proportion of unsuccessful downloads (SERVER ERROR or FORBIDDEN) are high in edu.py (15%) and edu.uy (17%) and can be attributed to broken links.

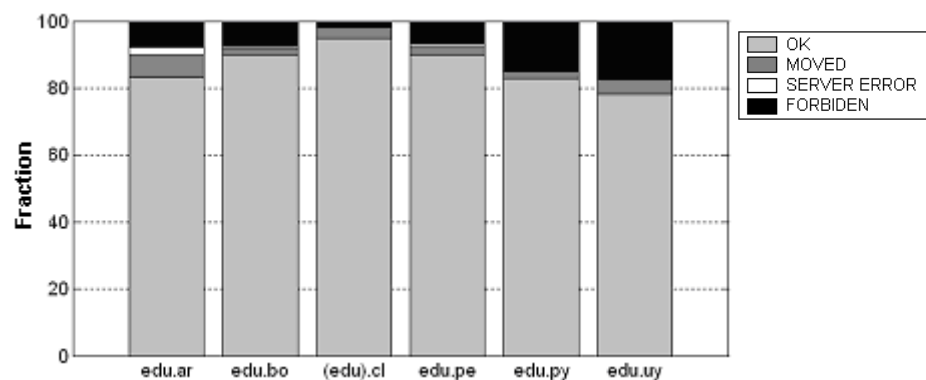


Figure 5 – Distribution of the most common HTTP codes

5.2 Page Depth

Another analyzed parameter was the logical depth of pages, which is the number of links that need to be followed in order to reach a page from the site's home page (depth 0). For this analysis, it should be recalled that the crawler was set to collect 15 levels for static pages and 5 for dynamic pages, as suggested in [5].

The majority of pages are at level 5 except edu.bo (level 6) and edu.py (level 3). This last case is another example of how recent this web space is. Up to level 6, we found between 70% and 90% of the pages, except in edu.ar, which accounts for 60% approximately.

5.3 Dynamic Pages

This section includes the distributions of links to dynamic pages, which gives a hint about the technologies used for their construction. As can be seen in Table 1 (Section 3), there exist between 12% and 30% of dynamic pages, which is why we decided to study the distribution of the programming tools used. It is easy to see that open source technologies predominate, especially

PHP³ (Figure 7). In Bolivia's educational web, we observed that this proportion is followed closely by the ASP technology.

This behavior is similar to what was observed in Chile [4], where PHP accounts for 78%, and Brazil, where PHP reaches 73%. In Spain [5], PHP represents 45% and ASP 42%.

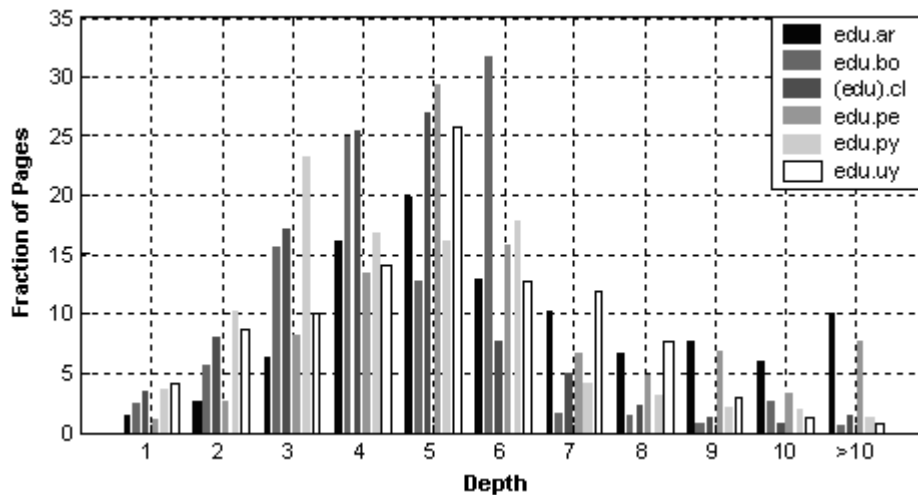


Figure 6 – Distribution of pages according to their depth.

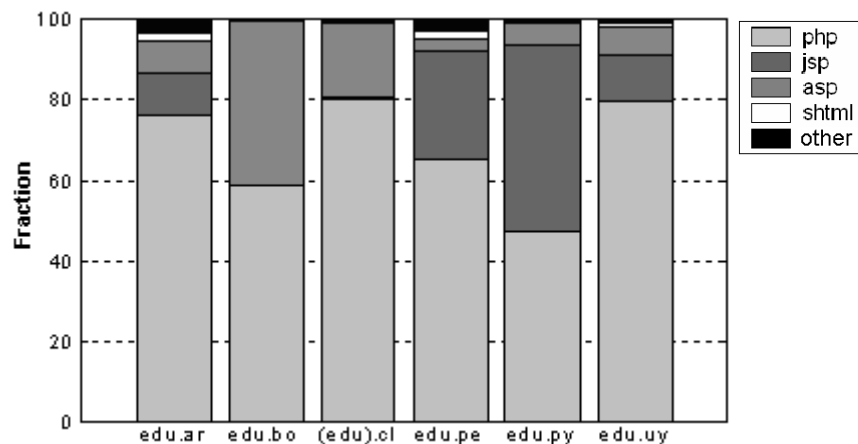


Figure 7 – Distribution of technologies used for dynamic pages.

5.4 Non-HTML Documents

Many links point to documents which are not in HTML format. All documents were withdrawn for further analysis. From the rest, more than 90%, in all cases, were links to image files (of different formats), which are generally needed for the design of web pages. Excluding these kind of links, the distribution of the remaining file types is shown in Table 7.

As for documents, we studied those which are most used (PDF, DOC, PS and TXT). PDF is the most used format, representing more than 60% of links. The PS format is only used in the

³ <http://www.php.net/>

educational webs of Argentina and Uruguay, with a smaller proportion in the latter. If we consider that this format is widely spread within the academic field, it becomes evident that it has lost popularity compared to PDF over the last years. This may seem consistent with the age of the studied webs and their relatively short life. This distribution is displayed in Figure 8.

		edu.ar	edu.bo	(edu).cl	edu.pe	edu.py	edu.uy
Type		%	%	%	%	%	%
Source Code	(c, h, cc, java, js, sh)	2.21	5.83	0.31	1.26	0.10	0.81
Video	(swf, mpg, avi, mov, qt)	19.52	64.99	84.34	80.29	78.27	71.19
Audio	(mp3, mid, wav, ram)	0.94	1.00	1.89	0.99	3.10	2.40
Software	(deb, exe, rpm, iso)	26.28	5.98	9.61	3.72	4.56	10.15
Compressed	(gz, tar, zip, rar)	51.05	22.20	3.85	13.74	13.97	15.45

Table 7 – Distribution of file types (not documents).

In Chile [4], we found a 63% of links to PDF documents and a 15% to DOC. In Brazil [16], PDF documents amount to 42% and 23% for DOC. Finally, a 41% of PDFs was observed in Spain [5] and, surprisingly, also a 41% of TXT documents, which barely represent an 8% in the examined domains (edu.ar).

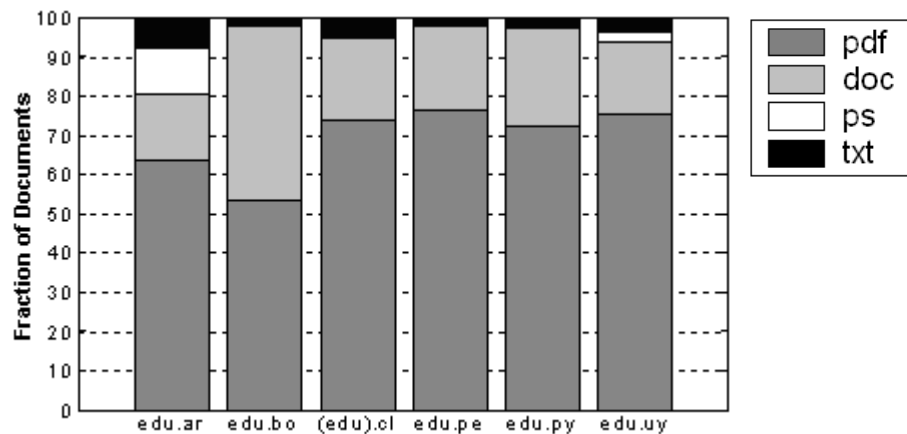


Figure 8 – Distribution of links to documents.

6 Conclusions and Future Work

As regards the growth of web spaces, we can state that the amount of contents in all the examined countries has increased considerably over the last year, whereupon their educational webs can still be considered young. The predominant language is Spanish, however, in Argentina and Bolivia there is a significant presence of English, with shares of 40% and 48% respectively.

Our study also reveals that it is necessary for pages in educational domains to gain more visibility among their peers. A 33% of pages with zero in-degree has been found in edu.ar, 54% in edu.bo, 20% in (edu).cl, 34% in edu.pe, 23% in edu.py and 39% in edu.uy. A good link structure would boost the importance of academic pages and improve their position on the result lists generated by search engines after submitting a query. Our study also showed that, besides HTML, PDF is the standard format for academic documents.

On the whole, new technologies are still needed to manage dynamic pages within educational institutions. The adoption of such technologies would extend the availability of academic contents and facilitate content publishing to the members of these communities. From the analysis of digital resources which are neither HTML nor documents, we found that within Argentina's web space there is a high percentage (77%) of links which point generally to software (compressed files and distribution packages) which is available to any web user.

The overall aim of this work is to provide people responsible for educational contents with revealing facts so they can adopt new tools to evaluate the state and evolution of their different national educational web spaces. Similar studies will be conducted in the future to monitor the development of the examined domains, trying to include new regional countries.

Acknowledgements

We would like to thank Ricardo Baeza-Yates for his invaluable comments and for providing us with the initial set of addresses of Chilean educational institutions. We are also grateful to Carlos Castillo for his assistance in many issues regarding the data collection process.

References

- [1] Albert R. and Barabasi A.-L. Statistical mechanics of complex networks. *Review of Modern Physics* 74, 47-94, 2002.
- [2] Baeza-Yates R. and Lalanne F. Characteristics of the Korean Web. Technical Report, KoreaChile IT Cooperation Center, ITCC, 2004.
- [3] Baeza-Yates R. and Castillo C. Crawling the infinite Web: five levels are enough. *Proceedings of the third Workshop on Web Graphs (WAW)*, Rome, Italy, Springer LNCS, October 2004.
- [4] Baeza-Yates R. and Castillo C. Características de la Web Chilena 2004. Technical Report, Center for Web Research, University of Chile, 2005.
- [5] Baeza-Yates R, Castillo C. and Lopez V. Characteristics of the Web of Spain. *Cybermetrics*, Vol. 9, No. 1, 2005.
- [6] Baeza-Yates R. and Castillo C. Link Analysis in National Web Domains. *Workshop on Open Source Web Information Retrieval (OSWIR)*, pp. 15-18. Compiègne, France, 2005.
- [7] Baeza-Yates R., Castillo C. and Efthimiadis E. Characterization of national Web domains. Technical Report, Universitat Pompeu Fabra, July 2005.
- [8] Barabasi A.-L. and Albert A. Emergence of Scaling in Random Networks. *Science*, (286): 509-512, 1999.

- [9] Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkings A. and Wiener J. Graph structure in the Web: Experiments and models. *Proceedings of the Ninth Conference on World Wide Web*, pp 309-320, Amsterdam, Netherlands, 2000.
- [10] Castillo C. and Baeza-Yates R.. WIRE: an Open-Source Web Information Retrieval Environment. *Workshop on Open Source Web Information Retrieval (OSWIR)*, pp. 27-30. Compiègne, France, September 2005.
- [11] Cho J. and García-Molina H. The evolution of the Web and implications for an incremental crawler. *Proceedings of the 26th International Conference on Very Large Data Bases*. 10–14, 200–209, September, 2000.
- [12] Dill S., Kumar R., Mccurley K.S., Rajagopalan S., Sivakumar D. and Tomkins A.. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205-223, 2002.
- [13] Efthimiadis E. and Castillo C. Charting the Greek Web. *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST)*, Providence, Rhode Island, USA, November, 2004.
- [14] Kleinberg J., Kumar R., Raghavan P., Rajagopalan S. and Tomkins A. The Web as a Graph: Measurements, Models and Methods. *Proceedings of the International Conference on Combinatorics and Computing*, 1999.
- [15] Kleinberg J. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632, 1999.
- [16] Modesto M., Pereira A., Ziviani N., Castillo C. and Baeza-Yates R.. Un novo retrato da Web Brasileira. *Proceedings of SEMISH*, Sao Leopoldo, Brazil, 2005.
- [17] O'Neill E., Lavoie B. and Bennett R. Trends in the Evolution of the Public Web 1998 - 2002. *D-Lib Magazine*, Vol. 9 Number 4, 2003.
- [18] Page L., Brin S., Montwani R and Winograd T.. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Library Technologies Project, 1998
- [19] Tolosa G. and Bordignon F. Análisis de Enlaces en el Espacio Web de las Universidades Argentinas. *Anales del VIII Workshop de Investigadores en Ciencias de la Computación (WICC 2006)*. ISBN 978-950-9474-35-2. 2006.